

(19) 日本国特許庁 (J P)

(12) 公 開 特 許 公 報 (A)

(11) 特許出願公開番号
特開2002-373059
(P2002-373059A)

(43) 公開日 平成14年12月26日 (2002. 12. 26)

(51) Int.Cl. ⁷	識別記号	F I	テーマコード (参考)
G 0 6 F 3/06	5 4 0	G 0 6 F 3/06	5 4 0 5 B 0 1 8
12/16	3 2 0	12/16	3 2 0 L 5 B 0 6 5

審査請求 未請求 請求項の数 7 O L (全 10 頁)

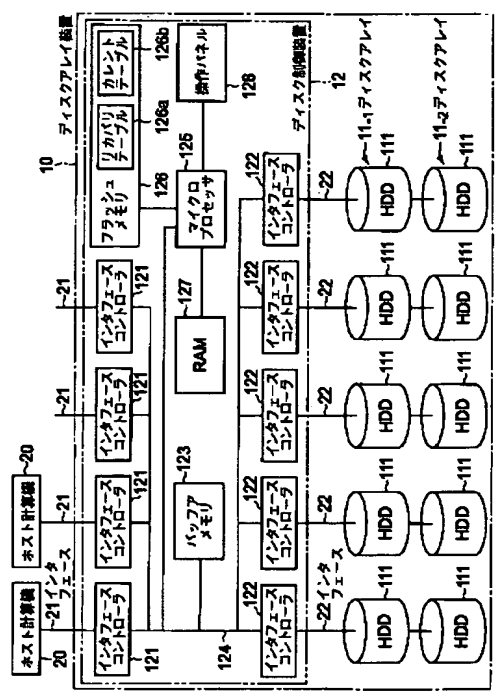
(21) 出願番号	特願2001-180202 (P2001-180202)	(71) 出願人	000003078 株式会社東芝 東京都港区芝浦一丁目1番1号
(22) 出願日	平成13年6月14日 (2001. 6. 14)	(72) 発明者	笹本 享一 東京都府中市東芝町1番地 株式会社東芝 府中事業所内
		(74) 代理人	100058479 弁理士 鈴江 武彦 (外6名)
		Fターム (参考)	5B018 GA06 KA15 MA14 5B065 EA11 EA18

(54) 【発明の名称】 ディスクアレイのエラー回復方法、ディスクアレイ制御装置及びディスクアレイ装置

(57) 【要約】

【課題】 ディスクドライブの多重故障によりディスクアレイが閉塞したとしても、当該ディスクアレイを簡単に使用可能な状態に復旧できるようにする。

【解決手段】 ディスクアレイ制御装置12内のマイクロプロセッサ125は、ディスクアレイ11-i内の複数のHDD111の故障により当該アレイが閉塞する際、閉塞直前のカレントテーブル126b上のRAID構成情報をリカバリテーブル126aに保存し、当該アレイの閉塞後に、カレントテーブル126b上のRAID構成情報を当該アレイの閉塞後の状態を反映するように更新する。ユーザが操作パネル128を操作することでリカバリ要求が与えられた場合、マイクロプロセッサ125はリカバリテーブル126aから指定のディスクアレイ11-iのRAID構成情報を読み出し、当該情報に従ってアレイ11-iを閉塞直前の状態に戻し、当該情報をカレントテーブル126bに上書きする。



(2) 002-373059 (P2002--腺穀

【特許請求の範囲】

【請求項1】 複数のディスクドライブから構成されるRAID (Redundant Arrays of Inexpensive Disks) 構成の少なくとも1つのディスクアレイを備え、と共に、現在の前記ディスクアレイの少なくとも構成及び稼働状態を示す第1のRAID構成情報が保存される不揮発性の第1の記憶領域と、前記ディスクアレイが閉塞した際に、その閉塞直前の当該ディスクアレイの少なくとも構成及び稼働状態を示す第2のRAID構成情報が保存される第2の記憶領域とが確保されたディスクアレイ装置におけるディスクアレイのエラー回復方法であって、

前記ディスクアレイを構成する前記複数のディスクドライブのうちの少なくとも2つのディスクドライブが故障した場合に、当該ディスクアレイを閉塞するステップと、

前記ディスクアレイを閉塞するに際し、前記第1の記憶領域に保存されている当該ディスクアレイの閉塞直前の第1のRAID構成情報を前記第2のRAID構成情報として前記第2の記憶領域に保存するステップと、

前記ディスクアレイの閉塞後、当該ディスクアレイの前記第1の記憶領域上の前記第1のRAID構成情報を当該ディスクアレイの閉塞後の状態を反映するように更新するステップと、

前記ディスクアレイ装置に対して閉塞状態にある前記ディスクアレイを使用可能な状態に回復させるリカバリ要求がユーザ操作に従って与えられた場合に、前記第2の記憶領域から、当該リカバリ要求で指定されたディスクアレイの閉塞直前の前記第2のRAID構成情報を読み出すステップと、

前記第2の記憶領域から読み出された前記第2のRAID構成情報に基づき前記ディスクアレイを閉塞直前の状態に強制的に戻すと共に当該第2のRAID構成情報を前記第1のRAID構成情報として前記第1の記憶領域に上書きするステップとを具備することを特徴とするディスクアレイのエラー回復方法。

【請求項2】 前記ディスクアレイを閉塞するに際し、前記第2のRAID構成情報と共に当該ディスクアレイが閉塞した原因を示す情報も前記第2の記憶領域に保存し、

前記リカバリ要求に従って前記ディスクアレイを閉塞直前の状態に強制的に戻す際には、前記第2の記憶領域から、当該ディスクアレイの閉塞直前の前記第2のRAID構成情報に加えて当該ディスクアレイが閉塞した原因を示す情報を読み出すステップと、

読み出された閉塞原因情報がディスクドライブの部分的な障害であるメディアエラーを示している場合、当該メディアエラーの発生していたブロックを代替用のセクタブロックに代替処理するステップとを更に具備し、前記代替処理後に、前記閉塞直前の前記ディスクアレイ

の前記第2のRAID構成情報を前記第1のRAID構成情報として前記第1の記憶領域に上書きすることを特徴とする請求項1記載のディスクアレイのエラー回復方法。

【請求項3】 複数のディスクドライブから構成されるRAID (Redundant Arrays of Inexpensive Disks) 構成の少なくとも1つのディスクアレイを制御するディスクアレイ制御装置において、

現在の前記ディスクアレイの少なくとも構成及び稼働状態を示す第1のRAID構成情報が保存される第1の記憶領域と、前記ディスクアレイが閉塞した際に、その閉塞直前の当該ディスクアレイの少なくとも構成及び稼働状態を示す第2のRAID構成情報が保存される第2の記憶領域とが確保された不揮発性記憶手段と、

前記ディスクアレイを構成する前記複数のディスクドライブのうちの少なくとも2つのディスクドライブが故障した場合に、当該ディスクアレイを閉塞する手段と、

前記閉塞手段により前記ディスクアレイが閉塞される際に、前記第1の記憶領域に保存されている当該ディスクアレイの閉塞直前の第1のRAID構成情報を前記第2のRAID構成情報として前記第2の記憶領域に保存すると共に、前記ディスクアレイの閉塞後、当該ディスクアレイの前記第1の記憶領域上の前記第1のRAID構成情報を当該ディスクアレイの閉塞後の状態を反映するように更新する閉塞時RAID構成情報更新手段と、

前記ディスクアレイ装置に対して閉塞状態にある前記ディスクアレイを使用可能な状態に回復させるリカバリ要求がユーザ操作に従って与えられた場合に、前記第2の記憶領域から、当該リカバリ要求で指定されたディスクアレイの閉塞直前の前記第2のRAID構成情報を読み出す手段と、

前記第2の記憶領域から読み出された前記第2のRAID構成情報に基づき前記ディスクアレイを閉塞直前の状態に強制的に戻すと共に当該第2のRAID構成情報を前記第1のRAID構成情報として前記第1の記憶領域に上書きする復旧手段とを具備することを特徴とするディスクアレイ制御装置。

【請求項4】 複数のディスクドライブから構成されるRAID (Redundant Arrays of Inexpensive Disks) 構成の少なくとも1つのディスクアレイであって、現在の当該ディスクアレイの少なくとも構成及び稼働状態を示す第1のRAID構成情報が保存される第1の記憶領域と、前記ディスクアレイが閉塞した際に、その閉塞直前の当該ディスクアレイの少なくとも構成及び稼働状態を示す第2のRAID構成情報が保存される第2の記憶領域とが確保された複数のディスクドライブから構成されるディスクアレイを制御するディスクアレイ制御装置であり、

前記ディスクアレイを構成する前記複数のディスクドライブのうちの少なくとも2つのディスクドライブが故障

(3) 002-373059 (P2002-M59)

した場合に、当該ディスクアレイを閉塞する手段と、前記閉塞手段により前記ディスクアレイが閉塞される際に、前記第1の記憶領域に保存されている当該ディスクアレイの閉塞直前の第1のRAID構成情報を前記第2のRAID構成情報として前記第2の記憶領域に保存すると共に、前記ディスクアレイの閉塞後、当該ディスクアレイの前記第1の記憶領域上の前記第1のRAID構成情報を当該ディスクアレイの閉塞後の状態を反映するように更新する閉塞時RAID構成情報更新手段と、前記ディスクアレイ装置に対して閉塞状態にある前記ディスクアレイを使用可能な状態に回復させるリカバリ要求がユーザ操作に従って与えられた場合に、前記第2の記憶領域から、当該リカバリ要求で指定されたディスクアレイの閉塞直前の前記第2のRAID構成情報を読み出す手段と、前記第2の記憶領域から読み出された前記第2のRAID構成情報に基づき前記ディスクアレイを閉塞直前の状態に強制的に戻すと共に当該第2のRAID構成情報を前記第1のRAID構成情報として前記第1の記憶領域に上書きする復旧手段とを具備することを特徴とするディスクアレイ制御装置。

【請求項5】 前記閉塞時RAID構成情報更新手段は、前記閉塞手段により前記ディスクアレイが閉塞される際に、前記第2のRAID構成情報と共に当該ディスクアレイが閉塞した原因を示す情報を前記第2の記憶領域に保存するように構成され、前記リカバリ要求が与えられた場合に前記読み出し手段によって前記第2の記憶領域から読み出された前記閉塞原因情報がディスクドライブの部分的な障害であるメディアエラーを示している場合、当該メディアエラーの発生していたブロックを代替用のセクタブロックに代替処理するセクタブロック代替手段を更に具備することを特徴とする請求項3または請求項4記載のディスクアレイ制御装置。

【請求項6】 複数のディスクドライブから構成されるRAID (Redundant Arrays of Inexpensive Disks) 構成の少なくとも1つのディスクアレイと、前記ディスクアレイを制御するディスクアレイ制御装置とを備えたディスクアレイ装置において、現在の前記ディスクアレイの少なくとも構成及び稼働状態を示す第1のRAID構成情報が保存される第1の記憶領域と、前記ディスクアレイが閉塞した際に、その閉塞直前の当該ディスクアレイの少なくとも構成及び稼働状態を示す第2のRAID構成情報が保存される第2の記憶領域とが確保された不揮発性記憶手段を備えると共に、前記ディスクアレイ制御装置は、前記ディスクアレイを構成する前記複数のディスクドライブのうちの少なくとも2つのディスクドライブが故障した場合に、当該ディスクアレイを閉塞する手段と、

前記閉塞手段により前記ディスクアレイが閉塞される際に、前記第1の記憶領域に保存されている当該ディスクアレイの閉塞直前の第1のRAID構成情報を前記第2のRAID構成情報として前記第2の記憶領域に保存すると共に、前記ディスクアレイの閉塞後、当該ディスクアレイの前記第1の記憶領域上の前記第1のRAID構成情報を当該ディスクアレイの閉塞後の状態を反映するように更新する閉塞時RAID構成情報更新手段と、前記ディスクアレイ装置に対して閉塞状態にある前記ディスクアレイを使用可能な状態に回復させるリカバリ要求がユーザ操作に従って与えられた場合に、前記第2の記憶領域から、当該リカバリ要求で指定されたディスクアレイの閉塞直前の前記第2のRAID構成情報を読み出す手段と、

前記第2の記憶領域から読み出された前記第2のRAID構成情報に基づき前記ディスクアレイを閉塞直前の状態に強制的に戻すと共に当該第2のRAID構成情報を前記第1のRAID構成情報として前記第1の記憶領域に上書きする復旧手段とを備えることを特徴とするディスクアレイ装置。

【請求項7】 前記閉塞時RAID構成情報更新手段は、前記閉塞手段により前記ディスクアレイが閉塞される際に、前記第2のRAID構成情報と共に当該ディスクアレイが閉塞した原因を示す情報を前記第2の記憶領域に保存するように構成され、前記ディスクアレイ制御装置は、前記リカバリ要求が与えられた場合に前記読み出し手段によって前記第2の記憶領域から読み出された前記閉塞原因情報がディスクドライブの部分的な障害であるメディアエラーを示している場合、当該メディアエラーの発生していたブロックを代替用のセクタブロックに代替処理するセクタブロック代替手段を更に備えていることを特徴とする請求項6記載のディスクアレイ装置。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】 本発明は、RAID (Redundant Arrays of Inexpensive Disks) 構成のディスクアレイ装置に係り、特にディスクアレイを構成するメンバーのディスクドライブが複数故障した場合に当該ディスクアレイを使用可能な状態に回復するのに好適なディスクアレイのエラー回復方法、ディスクアレイ制御装置及びディスクアレイ装置に関する。

【0002】

【従来の技術】 一般にディスクアレイ装置は、複数のディスクドライブ、例えば磁気ディスクドライブ（以下、HDDと称する）から構成される少なくとも1つのディスクアレイと、このディスクアレイ内の各HDD（メンバーHDD）に対するアクセスを制御するディスクアレイ制御装置とを備えている。

【0003】 ディスクアレイ装置は以下に述べる2つの

(4) 002-373059 (P2002-159)

特徴を有する外部記憶装置として知られている。第1の特徴は、ホスト計算機上のファイルシステムから要求されたデータの読み出し／書き込みを、ディスクアレイ内の各HDDを並列に動かして分散して実行することでアクセスの高速化を図っている点である。第2の特徴は、データの冗長化によって信頼性の向上を図っている点である。

【0004】ディスクアレイ制御装置は、ホスト計算機から転送される書き込みデータに対して、データ訂正情報としての冗長データを生成する。ディスクアレイ制御装置は、この冗長データをディスクアレイ内の複数のHDDのうちのいずれか1つに書き込む。これにより、複数のHDDのうちの1台が故障した場合、この冗長データと残りの正常なHDDのデータを用いて故障したHDDのデータを修復することが可能となる。

【0005】データ冗長化の手法の1つとして、RAID手法が知られている。RAID手法は、RAIDのデータと冗長データとの関連において、種々のRAIDレベルに分類される。RAIDレベルの代表的なものにレベル3とレベル5がある。

【0006】レベル3（RAIDレベル3）では、ディスクアレイ制御装置は、ホスト計算機から転送される更新データ（書き込みデータ）を分割して、その分割された更新データ間の排他的論理和演算を行うことで冗長データとしてのパリティデータを生成する。ディスクアレイ制御装置は、このパリティデータで複数のHDDのいずれかに書き込まれている元のパリティデータを更新する。一方、レベル5（RAIDレベル5）では、ディスクアレイ制御装置は、ホスト計算機から転送される更新データ（新データ）と、当該更新データの格納先となるHDD内領域に格納されている更新前のデータ（旧データ）と、当該更新データの格納先に対応する別のHDDの領域に格納されている更新前のパリティデータ（旧パリティデータ）との間の排他的論理和演算を行うことで、更新されたパリティデータ（新パリティデータ）を生成する。ディスクアレイ制御装置は、この新パリティデータで元のパリティデータを更新する。

【0007】このようなRAID構成のディスクアレイ装置では、ディスクアレイ内のメンバーHDDが故障した場合に、故障HDDのデータが次のように修復される。まずディスクアレイ制御装置は、故障したHDD以外の各HDDから、ディスクアレイのディスク領域を管理する単位であるストライプ毎にデータを読み出す。ディスクアレイ制御装置は、各HDDから読み出したデータの排他的論理和演算を行うことで、修復された（復元された）データを取得する。この排他的論理和演算を用いた手法、つまりRAIDのデータとパリティデータ（冗長データ）との整合性を利用したデータ修復を含む手法がRAID手法である。ディスクアレイ制御装置は、このRAID手法を用いて、修復されたデータをス

トライプ毎に取得することにより、故障したHDDのすべての領域のデータを、故障したHDDに代えて用いられるHDD内に修復することができる。この故障したHDDに代えて用いられるHDDは、故障したHDDと交換して用いられるHDD、またはディスク制御装置に予め接続されていて、故障したHDDの代替として割り付けられるスペアHDDである。

【0008】このように、RAID構成のディスクアレイ装置では、ディスクアレイ内のメンバーHDDが故障しても、故障したHDDのデータを元通りに修復することができる。

【0009】しかしながら、同一のディスクアレイ内で複数のメンバーHDDが故障する、いわゆるHDDの多重故障が発生した場合には、RAIDのデータ冗長性を利用してユーザデータを修復することはできない。つまり、HDDの多重故障が発生した場合、RAID手法によってユーザデータを修復することはできない。この場合、ディスクアレイ全体の故障となり、そのアレイは閉塞して、そのアレイ内のデータへはアクセスすることができなくなる。

【0010】このようにディスクアレイが閉塞した場合、従来は故障したすべてのHDDを交換し、改めてディスクアレイの状態を初期化した後に、別途データのバックアップが採取されたテープデバイスなどから、このアレイに対しデータを書き戻すのが一般的であった。

【0011】

【発明が解決しようとする課題】上記したように、RAID構成の従来のディスクアレイ装置では、ディスクアレイが閉塞した場合、故障したすべてのHDDを交換し、改めてディスクアレイの状態を初期化した後に、データのバックアップが採取されたテープデバイスなどから、このアレイに対しデータを書き戻す必要があった。

【0012】しかしながら、テープデバイスからの書き戻しには長時間を要するため、その間システムが停止状態となる問題があった。また、万一データのバックアップがなかった場合には、システムの復旧には膨大な時間と労力を必要とした。また、データのバックアップがあった場合でも、テープデバイスからの書き戻しには長時間を要するため、その間システムが停止状態となる問題もあった。

【0013】本発明は上記事情を考慮してなされたものでその目的は、同一ディスクアレイ内でのディスクドライブの多重故障により当該ディスクアレイが閉塞したとしても、HDD故障が一過性または部分的なものであったならば、外部からの要求に応じて当該ディスクアレイを簡単に使用可能な状態に復旧することができるディスクアレイのエラー回復方法、ディスクアレイ制御装置及びディスクアレイ装置を提供することにある。

【0014】

【課題を解決するための手段】本発明に係るディスクア

(5) 002-373059 (P2002-359)

レイのエラー回復方法は、複数のディスクドライブから構成されるRAID構成の少なくとも1つのディスクアレイを備えると共に、現在のディスクアレイの少なくとも構成及び稼働状態を示すRAID構成情報（第1のRAID構成情報）が保存される不揮発性のカレント用記憶領域（第1の記憶領域）と、ディスクアレイの閉塞時には、その閉塞直前の当該ディスクアレイのリカバリ用RAID構成情報（第2のRAID構成情報）が保存されるリカバリ用記憶領域（第2の記憶領域）とが確保されたディスクアレイ装置におけるディスクアレイのエラー回復方法であって、ディスクアレイを構成する複数のディスクドライブのうちの少なくとも2つのディスクドライブが故障した場合に、当該ディスクアレイを閉塞するに際し、カレント用記憶領域に保存されている当該ディスクアレイの閉塞直前のRAID構成情報をリカバリ用RAID構成情報としてリカバリ用記憶領域に保存すると共に、当該ディスクアレイの閉塞後に、当該ディスクアレイのカレント用記憶領域上のRAID構成情報を当該ディスクアレイの閉塞後の状態を反映するように更新し、閉塞状態にあるディスクアレイを使用可能な状態に回復させるリカバリ要求がユーザ操作により与えられた場合に、リカバリ用記憶領域から回復対象となるディスクアレイの閉塞直前のリカバリ用RAID構成情報を読み出して、当該RAID構成情報に基づき上記回復対象となるディスクアレイを閉塞直前の状態に強制的に戻すと共に当該リカバリ用RAID構成情報をカレント用記憶領域に上書きすることを特徴とする。

【0015】このように本発明においては、ディスクアレイ内の複数のディスクドライブの故障（つまりディスクドライブの多重故障）により当該アレイが閉塞する際、閉塞直前のRAID構成情報をリカバリ用RAID構成情報としてリカバリ用記憶領域に保存し、当該アレイの閉塞後に、カレント用記憶領域上のRAID構成情報を当該アレイの閉塞後の状態を反映するように更新する。そして、ユーザの操作により閉塞状態にあるディスクアレイを対象とするリカバリ要求が与えられた場合、上記リカバリ用RAID構成情報を読み出して当該RAID構成情報に従い、指定されたディスクアレイを閉塞直前の構成や稼働状態に強制的に戻し、また当該RAID構成情報をカレント用記憶領域に上書きする。

【0016】これにより、ディスクドライブの多重故障にて対応するディスクアレイが閉塞されていても、緊急的にそのアレイの稼働を再開したい、または重要なデータだけでもそのアレイ内から読み出してバックアップを取りたいなどの理由で、ユーザの操作によりリカバリ要求が与えられた場合、ディスクドライブの故障が当該ドライブの電源をOFF/ONしたり、モジュールを抜き差ししたりすることにより回復してしまう一過性のものであったならば、そのアレイを簡単に使用可能な状態に復旧することが可能となる。

【0017】さて、ディスクドライブの故障には、上記一過性のものの他に、一部分の領域でのみ発生する障害（部分的な障害）、例えばメディアエラーがある。ディスクドライブの故障がメディアエラーに関する故障の場合、そのメディアエラーの発生していたセクタブロックを代替用のセクタブロックに代替処理してから、上記した指定ディスクアレイを閉塞直前の構成や稼働状態に強制的に戻すリカバリ処理を行うとよい。そのためには、ディスクアレイを閉塞するに際し、RAID構成情報以外に当該ディスクアレイが閉塞した原因を示す情報もリカバリ用記憶領域に保存しておき、当該ディスクアレイのリカバリ処理を行う際にこの閉塞原因を示す情報も同時に読み出して、閉塞原因がメディアエラーにあるか否かを判定し、メディアエラーであるなら、上記のようにメディアエラーの発生していたセクタブロックを代替用のセクタブロックに代替処理してからリカバリ処理を行うとよい。

【0018】このようにすると、ディスクアレイ内の殆どのユーザデータは回復され、システムの稼働を継続することが可能となる。

【0019】

【発明の実施の形態】以下、本発明の実施の形態につき図面を参照して説明する。図1は本発明の一実施形態に係るディスクアレイ装置の構成を示すブロック図である。

【0020】図1において、ディスクアレイ装置10は、少なくとも1つのディスクアレイ、例えば2つのディスクアレイ11-1（#1）、11-2（#2）と、このディスクアレイ11-1（#1）、11-2（#2）を制御するディスクアレイ制御装置12とから構成される。

【0021】各ディスクアレイ11-1、11-2は、いずれも複数のHDD（磁気ディスクドライブ）、例えば5台のHDD（メンバーHDD）111から構成される。

【0022】ディスクアレイ制御装置12は、ディスクアレイ装置10を利用するホスト計算機20とインタフェース21を介して接続するための複数のインタフェースコントローラ121と、ディスクアレイ11-1、11-2を構成する各HDD111とインタフェース22を介して接続するための同数のインタフェースコントローラ122と、ホスト計算機20とディスクアレイ装置10との間で転送されるデータを一時的に格納するバッファメモリ123と、これらインタフェースコントローラ121及び122とバッファメモリ123との間のデータ転送に用いられるデータバス124とを備えている。インタフェース21及び22は、例えばSCSI（Small Computer System Interface）、或いはファイバチャネル（FibreChannel）である。バッファメモリ123は、ディスクアレイ11-1、11-2の一部のデータの写しが保持されるディスクキャッシュを含んでいる。データバス124は、例えばPCI（Peripheral Component Int

!(6) 002-373059 (P2002-ch>59

erconnect Bus) である。

【0023】ディスクアレイ制御装置12はまた、当該制御装置12の主制御部をなし、当該制御装置12全体とディスクアレイ11-1、11-2とを制御するマイクロプロセッサ125と、マイクロプロセッサ125が実行する制御プログラム等が格納される不揮発性メモリ、例えばフラッシュメモリ126と、マイクロプロセッサ125のワークエリア等を提供する揮発性メモリ、例えばRAM127と、ディスクアレイ制御装置12の状態表示及びディスクアレイ制御装置12に対するユーザ操作による指示入力等に用いられる操作パネル128とを備えている。

【0024】フラッシュメモリ126には、リカバリテーブル126aとカレントテーブル126bとの2つのテーブルが置かれる。リカバリテーブル126aはディスクアレイ11-iが閉塞した際に、後で当該アレイ11-iを使用可能な状態に戻す際の情報を保存するのに用いられる。この情報は、ディスクアレイ11-iについての閉塞直前のRAID構成情報とアレイ11-iが閉塞した原因を示す情報とから構成される。カレントテーブル126bは、現在のディスクアレイ11-iの状態を示すRAID構成情報を保存するのに用いられる。

【0025】ディスクアレイ11-iのRAID構成情報は、

- (1) ディスクアレイ11-iで適用されるRAIDレベル、ストライプサイズ、論理容量（ユーザに提供されるディスクアレイ11-i全体のディスク容量）を含むRAID基本情報
- (2) ディスクアレイ11-iを構成するHDD（メンバーHDD）111に関する情報（HDD番号及び台数を含む）
- (3) ディスクアレイ11-iの稼働状態（アレイ全体として正常であるか閉塞しているか）と各メンバーHDD111の稼働状態（正常/故障）とを示す情報から構成される。

【0026】ディスクアレイ制御装置12は、複数のHDD（メンバーHDD）111によって構成されるディスクアレイ11-i（ $i=1, 2$ ）のディスク領域を、図2に示すように複数のストライプ201に分割して管理する。

【0027】ストライプ201は、RAIDの手法により冗長性をもってストライピング配置された基本単位である。ストライピング配置とは、連続したデータを連続したディスクアレイ11-iに順次マッピングすることという。冗長性とは、ストライプ201に含まれる各HDD111の1つに、他の全HDD111のデータの排他的論理和値が冗長データとして格納されていることをいう。この冗長性により、ディスクアレイ11-1を構成する各HDD111の1つが故障しても、ストライプ201に含まれる他の全HDD111のデータの排他的論理

和をとることにより、故障したHDD111の修復されたデータを生成することができる。ストライプ201のサイズは、1HDD111当たり64K（キロ）バイト～256Kバイト程度に設定されるのが一般的である。

【0028】ストライプ201内のデータは、HDD111の最小アクセス単位であるセクタブロック202を単位にアクセスされる。セクタブロック202のサイズは512バイトであるのが一般的である。

【0029】各HDD111には、セクタブロック202がエラー（メディアエラー）となった場合、つまり不良セクタとなった場合に、当該不良セクタの代替用として用いられる専用のセクタブロック、いわゆる代替ブロック（代替セクタブロック）203が複数確保されている。

【0030】次に、図1の構成のディスクアレイ装置10における動作について、（A）ホスト計算機20からのデータアクセス要求受信時の動作、（B）HDD故障検出時の動作、（C）ユーザ操作によるディスクアレイのリカバリ（復旧）要求の受信時の動作を例に、順次説明する。

【0031】（A）データアクセス要求受信時の動作
まず、ホスト計算機20からディスクアレイ装置10に対してデータアクセス要求が発行された場合の当該ディスクアレイ装置10の動作を、図3のフローチャートを参照して説明する。

【0032】ホスト計算機20からインタフェース21を介してディスクアレイ装置10に発行されたデータアクセス要求は、ディスクアレイ制御装置12内のインタフェースコントローラ121で受け取られて、当該ディスクアレイ制御装置12内のマイクロプロセッサ125に渡される。

【0033】マイクロプロセッサ125は、ホスト計算機20からのデータアクセス要求を受け取ると、アクセスの対象となるディスクアレイ（対応ディスクアレイ）11-i（ i は1または2）が稼働中であるか或いは閉塞中であるかを判定する（ステップS1）。この判定は、フラッシュメモリ126上に確保されたカレントテーブル126bに保存されている対応ディスクアレイ11-iのRAID構成情報（に含まれている当該ディスクアレイ11-iの稼働状態を示す情報）に基づいて行われる。

【0034】もし、対応ディスクアレイ11-iが稼働中の場合、マイクロプロセッサ125は当該ディスクアレイ11-iへのデータアクセス処理を起動する（ステップS2）。これに対し、対応ディスクアレイ11-iが閉塞中の場合には、マイクロプロセッサ125は当該ディスクアレイ11-iに対するアクセスを行わずに、エラー終了する（ステップS3）。その理由は、次の通りである。

【0035】まず、同一ディスクアレイ11-i内で複数のHDD（メンバーHDD）111が故障するHDDの

(7) 002-373059 (P2002-0059)

多重故障が発生した場合、RAIDの冗長性を利用したデータ修復が不可能となるため、当該ディスクアレイ11-iは閉塞する。もし、HDDの多重故障が発生したディスクアレイ11-iを閉塞せずに、使用可能な状態に継続すると、故障したHDDへのアクセスにより他の正常なHDDへのアクセスに悪影響（HDDのインタフェースがロックするなど）を与える可能性がある。そのため、HDDの多重故障が発生したディスクアレイ11-iは閉塞し、つまり他の正常なHDDが存在しても、ディスクアレイ11-i全体としては故障であるとして、以降アクセスしないようにしている。

【0036】(B) HDD故障検出時の動作

次に、ディスクアレイ11-iを構成するHDD111の故障を検出した場合の動作を、図4のフローチャートを参照して説明する。

【0037】マイクロプロセッサ125は、フラッシュメモリ126に格納されている制御プログラムに従い、各ディスクアレイ11-i中の各HDD111を定期的に監視する。ここでは、各HDD111の記憶内容を読み出すことにより当該HDD111の部分的な障害（メディアエラー）を検出するメディア検査処理が、閉塞中のディスクアレイ11-iを構成するHDD111も対象として行われる。

【0038】さて、上記メディア検査処理、或いは先のデータアクセス処理等で新たにHDD111の故障が検出された場合、マイクロプロセッサ125は、検出されたHDD111をメンバーHDDとする対応するディスクアレイ11-iが稼働中であるか或いは閉塞中であるかを判定する（ステップS11）。

【0039】もし、対応ディスクアレイ11-iが稼働中である場合、マイクロプロセッサ125は新たにメンバーHDD111の故障が検出されても、当該ディスクアレイ11-iはまだ稼働可能であるか否かを判定する（ステップS12）。ここでは、対応ディスクアレイ11-i内の他の全てのHDD111が正常である場合に、当該ディスクアレイ11-iは稼働可能と判定される。これに対し、対応ディスクアレイ11-i内の他の少なくとも1つのHDD111が既に故障である場合、つまり新たなHDD111の故障により、RAIDの冗長性をもってしても故障したHDD111データが修復不可能な多重の故障状態となった場合、当該ディスクアレイ11-iは稼働不可能であり、当該ディスクアレイ11-iを閉塞する必要があると判定される。

【0040】マイクロプロセッサ125は、対応ディスクアレイ11-iが稼働不可能であると判定した場合、当該ディスクアレイ11-iを閉塞するに際し、後で当該ディスクアレイ11-iの状態を使用可能な状態に戻す際の情報として、現在の、つまり閉塞直前のRAID構成情報と閉塞原因とを、フラッシュメモリ126上のリカバリテーブル126aに保存する（ステップS13）。

【0041】次にマイクロプロセッサ125は、対応ディスクアレイ11-iを閉塞する処理を行う（ステップS14）。このときマイクロプロセッサ125は、ディスクアレイ11-i内の複数のHDD111の故障により、RAIDのデータ冗長性をもってしても故障したHDD111内のデータを修復することが不可能となり、その結果当該ディスクアレイ11-iを閉塞したことを、操作パネル128またはホスト計算機20上の専用ソフトウェアを通して通知する。そしてマイクロプロセッサ125は、フラッシュメモリ126上のカレントテーブル126bに現在保存されている対応ディスクアレイ11-iのRAID構成情報を、ステップS14でのアレイ閉塞を反映した新たなRAID構成情報に更新し（ステップS15）、しかる後にステップS16に進む。

【0042】一方、ステップS11で対応ディスクアレイ11-iが閉塞中であると判定された場合、或いはステップS12でディスクアレイ11-iが稼働可能であると判定された場合、マイクロプロセッサ125はそのままステップS16に進む。

【0043】マイクロプロセッサ125はステップS16において、上記新たに故障が検出されたHDD111を閉塞する（ディスクアレイ11-iから切り離す）処理を行う。そしてマイクロプロセッサ125は、カレントテーブル126bに現在保存されている対応ディスクアレイ11-iのRAID構成情報を、ステップS16での故障HDD111の閉塞を反映した新たなRAID構成情報に更新する（ステップS17）。

【0044】(C) ディスクアレイのリカバリ要求受信時の動作

次に、閉塞中のディスクアレイ11-iを対象とするリカバリ（回復）処理が要求された場合の動作について、図5のフローチャートを参照して説明する。

【0045】ユーザにとって、HDDの多重故障にてディスクアレイが閉塞した場合においても、緊急的にそのアレイの稼働を再開したい、または重要なデータだけでもそのアレイ内から読み出してバックアップを取りたいことがある。このような場合、本実施形態ではディスクアレイ装置10のディスクアレイ制御装置12に設けられた操作パネル128をユーザが操作することで、所望のディスクアレイ11-iのリカバリを要求することが可能になっている。また本実施形態では、ディスクアレイ装置10を利用するホスト計算機20からも、当該ホスト計算機20にインストールされた専用のソフトウェアに従い、ユーザの操作に応じてホスト計算機20からディスクアレイ装置10のディスクアレイ制御装置12に対してディスクアレイ11-iのリカバリを要求することが可能になっている。

【0046】さて、ユーザ操作により操作パネル128またはホスト計算機20から発行されたディスクアレイリカバリ（復旧）要求はディスクアレイ制御装置12の

(8) 002-373059 (P2002-梅59)

マイクロプロセッサ125で受け付けられる。マイクロプロセッサ125は、このリカバリ要求を受け付けると、当該要求で指定されたディスクアレイ（対応ディスクアレイ）11-iが閉塞中であるか否かを、カレントテーブル126b上の該当するRAID構成情報に基づいて判定する（ステップS21）。

【0047】もし、対応ディスクアレイ11-iが閉塞中でないならば、マイクロプロセッサ125はそのままリカバリ要求に対するリカバリ処理を終了する。

【0048】これに対し、対応ディスクアレイ11-iが閉塞中であるならば、マイクロプロセッサ125はリカバリテーブル126aから、対応ディスクアレイ11-iが閉塞する直前のRAID構成情報及び閉塞した原因を読み出す（ステップS22）。そしてマイクロプロセッサ125は、閉塞の原因が、対応ディスクアレイ11-iの閉塞直前に故障となったHDD111（つまり対応ディスクアレイ11-iを構成するHDD111のうち、閉塞のトリガとなったHDD111）におけるセクタブロックのメディアエラー（部分的な障害）にあるか否かを判定する（ステップS23）。

【0049】もし、対応ディスクアレイ11-iの閉塞の原因が、セクタブロックのメディアエラーにあった場合、マイクロプロセッサ125は、当該セクタブロックのデータの読み出しテストを行う（ステップS24）。そしてマイクロプロセッサ125は、この読み出しテストにより、依然として読み出しが不可能であるか否か、つまりメディアエラーが再現するか否かを判定する（ステップS25）。

【0050】もし、ディスクアレイ11-iが閉塞する原因となったセクタブロックのメディアエラーが再現した場合、マイクロプロセッサ125は当該メディアエラーが再現したセクタブロックのデータを破棄し、当該セクタブロックを代替ブロックへ代替する代替処理を行う（ステップS26）。この代替処理により、メディアエラーが再現したセクタブロックへのアクセスは正しく行われる。但し、以前のデータは消失している。そこでマイクロプロセッサ125は、セクタブロックのメディアエラーにより部分的なデータ消失が発生したことをユーザに通知する（ステップS27）。この通知には、操作パネル128を通して通知する方法、或いはホスト計算機20上の上記専用ソフトウェアを通して通知する方法が適用可能である。

【0051】マイクロプロセッサ125はステップS27を実行すると、先にステップS22でリカバリテーブル126aから読み出した閉塞直前のRAID構成情報を、カレントテーブル126bに上書きし（ステップS28）、以降このRAID構成情報に従って動作する。

【0052】一方、上記ステップS23で対応ディスクアレイ11-iの閉塞原因がメディアエラー以外にあると判定された場合、或いは上記ステップS25で当該メ

ディアエラーが再現しないと判定された場合、マイクロプロセッサ125はそのまま上記ステップS28に進んで、ステップS22で読み出した閉塞直前のRAID構成情報を、カレントテーブル126bに上書きする。

【0053】以上により、一度閉塞したディスクアレイ11-iが、閉塞直前の使用できていた状態に復旧される。これによりディスクアレイ11-iが再度使用可能な状態となるため、緊急的に運用稼働を継続したり、そのディスクアレイ11-i内の重要なデータのバックアップを採取したりすることが可能となる。但し、このような状態に真に復旧できるのは、あくまで閉塞の原因となったHDD111の故障が一過性の場合や、メディアエラーのように部分的な障害の場合のみである。恒久的なHDD111故障の場合は、閉塞直前のRAID構成情報を用いて対応ディスクアレイ11-iを閉塞直前の状態に戻しても、原因となったHDD111の故障が取り除かれないため、当該HDD111の故障が再び検出される。この場合、図4のフローチャートに従う処理により、対応ディスクアレイ11-iは直ちに再度閉塞される。

【0054】以上に述べた実施形態では、リカバリテーブル126a及びカレントテーブル126bがディスクアレイ制御装置12に実装されたフラッシュメモリ126（書き換え可能な不揮発性メモリ）上に確保されているものとして説明したが、これに限るものではない。例えば、HDD111内の一部領域を、ユーザデータの格納用とは別に、リカバリテーブル126a及びカレントテーブル126bの領域を含む、ディスクアレイ装置10自身の管理情報の記憶用領域として割り当てるようにしてもよい。

【0055】なお、本発明は、上記実施形態に限定されるものではなく、実施段階ではその要旨を逸脱しない範囲で種々に変形することが可能である。更に、上記実施形態には種々の段階の発明が含まれており、開示される複数の構成要件における適宜な組み合わせにより種々の発明が抽出され得る。例えば、実施形態に示される全構成要件から幾つかの構成要件が削除されても、発明が解決しようとする課題の欄で述べた課題が解決でき、発明の効果の欄で述べられている効果が得られる場合には、この構成要件が削除された構成が発明として抽出され得る。

【0056】

【発明の効果】以上詳述したように本発明によれば、ディスクドライブの多重故障により当該ディスクドライブをメンバーとするディスクアレイが閉塞してアクセス不能な状態となった場合においても、当該ディスクアレイを閉塞する際に、当該ディスクアレイについての閉塞直前のRAID構成情報をリカバリ用RAID構成情報として不揮発性のリカバリ用記憶領域に保存しておくことにより、その後、閉塞したディスクアレイのリカバリ

(9) 002-373059 (P2002-0.159)

(復旧)がユーザ操作に従って要求された場合に、リカバリ用記憶領域に保存されているリカバリ用RAID構成情報に基づいて当該ディスクアレイを閉塞直前の状態に簡単に復旧することができる。これにより、ディスクアレイの閉塞を招いたディスクドライブの故障が一過性のものまたはメディアエラーのような部分的なものであった場合には、上記復旧後のディスクアレイをアクセスして緊急的に運用を継続したり、当該アレイ内の重要なデータのバックアップを採取することができ、ユーザデータの保護を図ると共にシステムに与える被害を最小限に抑えることができる。

【図面の簡単な説明】

【図1】本発明の一実施形態に係るディスクアレイ装置の構成を示すブロック図。

【図2】図1中のディスクアレイ11-i (i=1, 2)のディスク領域を管理するのに用いられるストライプ、HDD 111の最小単位であるセクタブロック及び当該セクタブロックが不良セクタとなった場合に、当該不良

セクタの代替用として用いられる代替ブロックの関係を説明するための図。

【図3】データアクセス要求受信時の処理手順を説明するためのフローチャート。

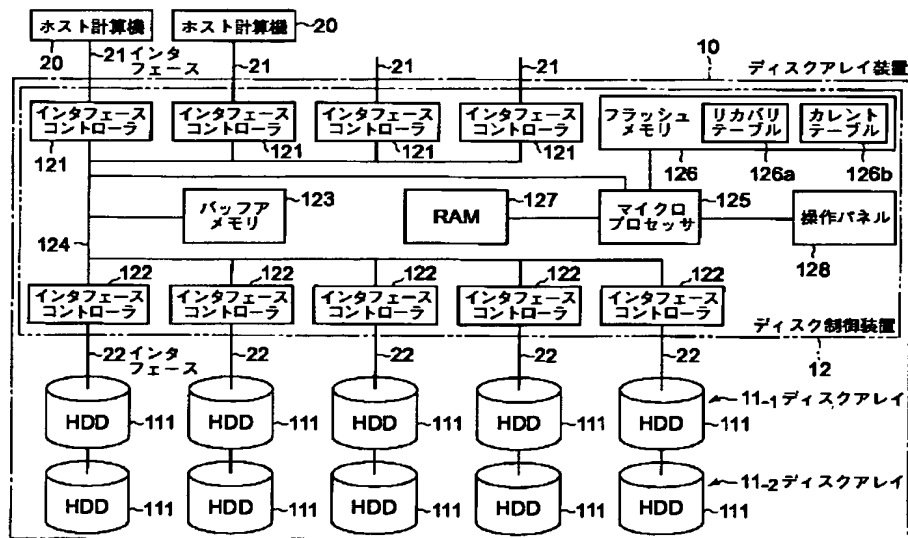
【図4】HDD故障検出時の処理手順を説明するためのフローチャート。

【図5】リカバリ要求受信時の処理手順を説明するためのフローチャート。

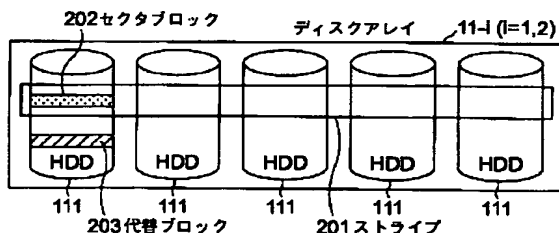
【符号の説明】

- 10…ディスクアレイ装置
- 11-1, 11-2, 11-i…ディスクアレイ
- 12…ディスクアレイ制御装置
- 111…HDD (ディスクドライブ)
- 125…マイクロプロセッサ
- 126…フラッシュメモリ
- 126a…リカバリテーブル (第2の記憶領域)
- 126b…カレントテーブル (第1の記憶領域)
- 128…操作パネル

【図1】

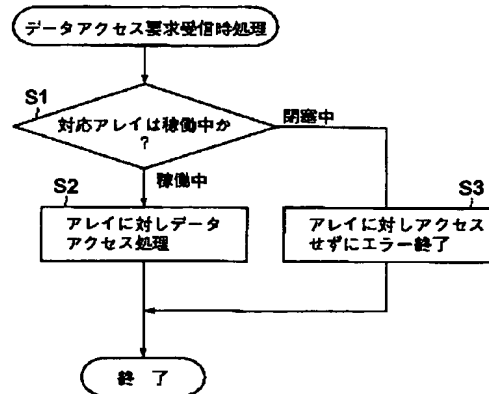


【図2】

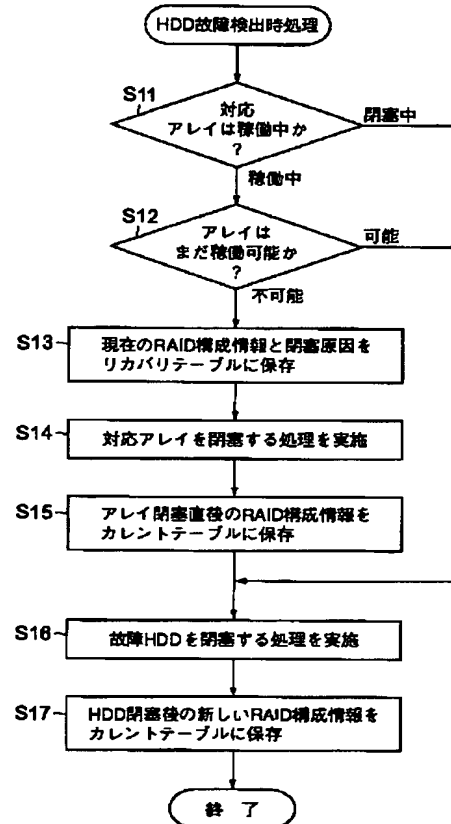


(株) 02-373059 (P2002-359)

【図3】



【図4】



【図5】

